# SAGE: Structure-Aware Generative Enhancement for Long-Tail Knowledge Graph Completion

Erzhuo Xu[1[0009-0003-2796-2955]] , Jun Huang[2*[0000-0001-8292-9160]] and Hailong Zhang[1[0009-0007-3559-846X]]

[1] Guangzhou University of Software, Guangzhou, 510990, China
[2] Guangxi Zhuang Autonomous Region Meteorological Bureau, Nanning City, 530011, China
874885365@qq.com

**Abstract.** Knowledge Graph Completion (KGC) remains a fundamental challenge, aiming to deduce missing links by reasoning over observed facts. While end-to-end deep learning has become the dominant paradigm for KGC, contemporary models are frequently bottlenecked by structural sparsity and suboptimal textual metadata. To mitigate these limitations, we propose SAGE, a novel framework that leverages Large Language Models (LLMs) for targeted data augmentation. Specifically, we first design a neuro-symbolic pre-extractor—integrating logical rules with neural networks—to identify long-tail entities and retrieve plausible candidate tails based on relational constraints. Subsequently, an LLM is prompted to evaluate these candidates and deduce the most factually accurate triplets. The resulting augmented graph is then utilized for downstream predictive inference. Empirical evaluations across three standard benchmarks demonstrate the superiority of SAGE; notably, on FB15k-237, it outperforms the competitive SimKGC baseline with absolute improvements of 1.0%, 0.9%, and 1.6% on Hits@1, Hits@3, and Hits@10, respectively.

**Keywords:** Knowledge Graph Completion, Data Augmentation, Large Models.

## 1 Introduction

By articulating entities and their interrelations, Knowledge Graphs (KGs) provide a structured repository of facts that underpins various downstream AI tasks, including question answering [2], recommendation systems [20], and semantic search. Since real-world KGs are inherently incomplete, Knowledge Graph Completion (KGC) has emerged as an essential mechanism to deduce unobserved connections, as illustrated in Fig. 1. Contemporary KGC methodologies are predominantly bifurcated into two paradigms: Knowledge Graph Embedding (KGE) techniques (e.g., TransE, RotatE), which project graph components into low-dimensional continuous spaces, and rule-based models (e.g., AMIE [8], TripleRE [26]), which leverage explicit logical patterns to perform highly interpretable reasoning. Despite their successes, these methods encounter a significant bottleneck when handling long-tail (or few-shot)

entities, which are associated with very few triples in real-world KGs. Conventional KGE models often learn poor representations for such sparse entities due to insufficient structural context, while rule-based methods frequently fail to fire as the prerequisite facts for logical inference are missing. Consequently, the performance of existing models deteriorates drastically in these data-sparse regions.

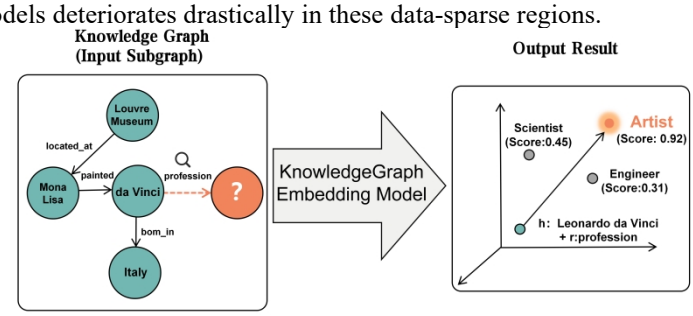


**Fig. 1.** Illustration of the knowledge graph completion process, where a knowledge graph embedding model predicts missing entities by scoring candidate tails for a given head-relation query.

To mitigate sparsity, recent studies have explored leveraging Large Language Models (LLMs)[11] to augment KGs by tapping into their rich parametric knowledge. Methods like MPIKGC[23] utilize prompting strategies such as Chain-of-Thought to enable LLM reasoning over KGs. However, directly employing LLMs for KGC introduces two prominent challenges: (1) LLMs may generate plausible yet factually incorrect triples, particularly when lacking precise contextual constraints; and (2) In multi-step reasoning processes, errors can propagate without retrieval grounding, steering the inference away from factual knowledge.

In response, we propose SAGE, a neuro-symbolic framework designed to retrieve long-tail entities and leverage LLMs for semantically-grounded augmentation, thereby strengthening KGC models. It begins with a structure-aware pre-selection mechanism that unifies logical rule reasoning with structural embeddings to produce a high-confidence candidate set, effectively pruning the search space. This step reframes the role of the frozen LLM from open-ended generation to constrained reasoning and verification in the subsequent semantic augmentation stage. By generating enriched triples and reasoning paths, SAGE constructs an augmented graph that alleviates the structural sparsity of long-tail entities. Finally, the model inference stage performs link prediction over this enriched graph, ensuring both structural consistency and semantic depth. The main contributions are summarized as follows:

- We propose SAGE, a structure-aware neuro-symbolic framework that leverages symbolic constraints to guide large language model – based data augmentation, effectively alleviating structural sparsity in long-tail entities while reducing hallucinations during generative inference.
- We design a logic-enhanced pre-extractor that integrates symbolic relational rules with neural embeddings to identify long-tail entities and generate

structurally consistent candidate triples, providing reliable and explainable inputs for subsequent LLM-based verification.

- Extensive experiments on three standard knowledge graph completion benchmarks demonstrate that SAGE consistently improves inference accuracy, with particularly notable gains on long-tail entities, outperforming strong baselines across multiple Hits@K metrics.

## 2 Related work

### 2.1 Knowledge Graph Reasoning and Deep Learning Approaches

Knowledge Graph Reasoning (KGR) serves as a fundamental task in knowledge graph research[25], primarily focusing on revealing implicit relationships[18], completing missing information, and verifying knowledge validity via logical reasoning or deep learning paradigms[16]. Existing deep learning-based knowledge graph completion (KGC) methods largely rely on logical rules or embedding-based techniques to predict missing links. To capture complex relational patterns (e.g., 1-to-N, N-to-N), Tang et al. extended relational representations from 2D complex space to high-dimensional space using Orthogonal Transformation Embedding (OTE), incorporating graph context to enhance prediction accuracy. More recently, hybrid frameworks have emerged to enforce stronger constraints: LSTK[14] imposes structural and textual knowledge constraints on rule-based learning, while NCRL[6] employs a recursive path-merging strategy with attention mechanisms to predict single predicates. Additionally, CompoundE3D[9] leverages 3D composite geometric transformations and breadth-first search to identify optimal model variants. Despite these advancements in model architecture, these methods often overlook the critical impact of data quality and integrity on reasoning performance.

### 2.2 Few-Shot Learning for Data Scarcity

A significant bottleneck in traditional deep learning models is their dependence on large-scale labeled data, which is often resource-intensive to acquire[3]. To mitigate this issue, researchers have investigated few-shot learning strategies that facilitate effective inference with minimal samples. For instance, matching networks have been utilized to aggregate representations from limited reference instances[27]. Furthermore, Niu et al. proposed a framework utilizing a gating and attention-based neighbor aggregator to filter noise from sparse neighborhoods[13]. By integrating meta-learning with TransH[22], this approach effectively models complex relationship patterns in low-resource settings. Despite their effectiveness, existing few-shot knowledge graph completion methods are primarily constrained by their dependence on local structural signals and predefined embedding spaces. In highly sparse or long-tail scenarios, the lack of informative neighbors and explicit relational semantics often limits their generalization ab

ility. Moreover, these approaches typically rely on purely discriminative learnin g objectives, which restrict their capacity to introduce novel but plausible facts beyond the observed graph structure. As a result, they struggle to fully exploi t external semantic knowledge or to perform structure-aware reasoning when fa ced with severe data scarcity.

## 3 Method

Here, we introduce a synergistic inference paradigm designed to bridge conti nuous graph representations with LLMs. A schematic overview of this methodo logy is provided in Fig. 2, illustrating how the process unfolds across two pri mary operational stages. (1) Neural Pre-extractor for Sparse Entity Expansion: To address entity sparsity in low-resource settings, we introduce a neural pre-e xtractor that operates on the original knowledge graph. This module jointly em beds entities, relations, and logical rules in a continuous vector space[17], enab ling it to capture both semantic and structural dependencies. A scoring function is then applied to rank and select contextually relevant candidate entities, ther eby dynamically expanding long-tail entities while preserving graph consistency. (2)LLM-based Knowledge Fusion: In the second stage, the high-confidence ca ndidates produced by the pre-extractor are organized into structured prompt tem plates and injected into an LLM[5]. Leveraging the implicit knowledge encoded inthe LLM, the model infers supplementary triples that are coherent with the contextual information of the original graph. These generated triples provide a controllable form of knowledge augmentation for few-shot entities [24]. To eva luate the effectiveness of the proposed enhancement strategy, the LLM-augment ed triples are incorporated into the training data and assessed using multiple k nowledge graph reasoning models.

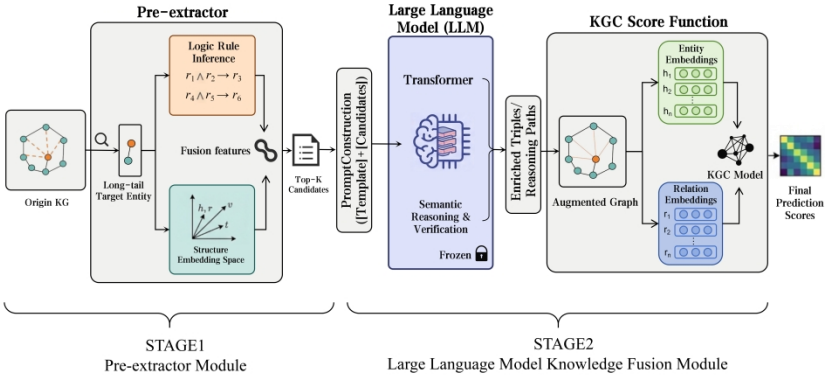


**Fig. 2**. The overall architecture of the SAGE framework designed to address the long-tail entity sparsity issue in Knowledge Graphs. The framework operates in a two-stage pipeline: (1) Pre-extractor Module, which retrieves top-K candidates by fusing logic rule inference with structural embeddings; (2) Large Language Model Knowledge Fusion Module, where high-

confidence candidates are formulated into prompts and verified by a frozen LLM to generate enriched triples.

### 3.1 Neural Pre-extractor for Sparse Entity Expansion

This section presents the neural pre-extractor designed to address entity sparsity in low-resource and long-tail scenarios. The core motivation of this module is to provide structure-aware candidate expansion that complements large language model–based reasoning, while avoiding uncontrolled or noisy augmentation. Instead of relying solely on local graph neighborhoods or textual semantics, the pre-extractor explicitly incorporates both structural relations and logical dependencies to generate reliable candidate entities. Consequently, our strategy bolsters KG inference capabilities through a holistic representation paradigm, which simultaneously maps entities, relations, and symbolic rules into a shared vector space. By integrating semantic representations with rule-based relational constraints, the pre-extractor serves as a semantic anchor that bridges sparse graph structures and downstream generative inference. This design enables the model to selectively expand long-tail entities with contextually consistent candidates, rather than indiscriminately increasing graph density. The pre-extractor consists of four key components, each responsible for capturing a complementary aspect of knowledge graph structure and semantics.

**Identification of Few-shot Entities**: In the study of entity distribution within knowledge graphs, the identification and handling of few-shot entities represent a critical step in improving the quality of the graph. This paper proposes a frequency threshold-based method for identifying few-shot entities. Initially, the pre-extractor traverses all entities in the knowledge graph using an internal entity counter to tally the occurrence frequency of each entity. This process employs a breadth-first traversal algorithm to ensure the completeness of the statistical results. Based on this, the system calculates the average occurrence frequency of entities using Equation (1):

$$\mu = \frac{\sum f\_e}{N} \tag{1}$$

Among them, f_e represents the frequency of occurrence of a single entity, and N is the total number of entities. Then, through triplet association analysis, a set is constructed for all the few-shot data contained in the knowledge grap h, providing structured data support for subsequent enhancement processing.

**Modeling Entities and Relationships in Triplets:** By seamlessly incorporati ng pre-existing symbolic rules directly into the representation space, RulE [17] significantly bolsters the generalization performance of traditional KGE framew orks. Due to the combination of rule-based and embedded knowledge graph co mpletion models, this paper uses RulE to model the few-shot data collected fro m the aforementioned few-shot entity recognition module. Specifically, for few-shot triplets (h, r, t), the scoring function is defined as follows:

$$S_t(h,r,t) = \gamma_t - d(h,r,t) \tag{2}$$

In this context, let $\gamma_t$ represent the predetermined margin. The fundamental premise of RotatE is that a true triplet (h,r,t) should satisfy t(c) ≈ (h(c)°r(c)) i ndicates the Hadamard product. This translates the prediction of the tail entity into a concatenation of two parts: the real part ha◦cos(r)-hb◦sin(r) and the imag inary part ha◦sin(r)+hb◦cos(r). Accordingly, the distance function evaluates the norm of the residual between this transformed head representation and the actu al tail embedding t. Triplets exhibiting smaller distances are deemed more fact ual and are assigned higher scores. Based on this formulation, we construct the loss function with negative sampling as:

$$L(R,\mathrm{r}_1,\mathrm{r}_2,...,\mathrm{r}_n) = -\log\sigma(\gamma_r - d(R,\mathrm{r}_1,\mathrm{r}_2,...,\mathrm{r}_n)) - \sum_{(h',f,t')} \frac{1}{N}\log\sigma(-s_t(h',r,t')) \tag{3}$$

In this formulation, the sigmoid function is represented by $\sigma$. To construct the negative candidate set N, we introduce random entity substitutions to either the head or the tail position of a true triplet, explicitly preventing simultaneou s corruption. Furthermore, inspired by RotatE, our approach leverages a self-ad versarial negative sampling paradigm to extract challenging negative triplets bas ed on the current state of the embedding model.

**Modeling Relations and Logical Rules** :Inspired by knowledge graph reaso ning based on logical rules[28], for logical rules: R: r_1 ∧ r_2 ∧ ... ∧ $\mathrm{r_l} \rightarrow \mathrm{r_{l+1}}$, we aim to approximate $r_{l+1}^{(c)} \approx (r_1^{(c)} \circ \cos r_2^{(c)} \circ \cos ... \circ \cos r_l^{(c)}) \circ \cos R^{(c)}$ , wh ere $\{r_i^{(c)}\}_i^l$ =1 and $\mathrm{R^c}$ are in K-dimensional complex space.A fundamental limi tation of two-dimensional rotations is their inherent commutativity, which precl udes them from accurately capturing the strictly non-commutative nature of co mpositional rules. For instance, consider the logical implication: sister of (x, y) ∧ mother of (y, z) → aunt of (x, z). Permuting the relations in the premis e to mother_of followed by sister_of fundamentally alters the semantic validity. However, rotation-based models erroneously assign identical scores to both seq uences due to their commutative property $(r_1^{(c)} \circ r_2^{(c)}) = (r_2^{(c)} \circ r_1^{(c)})$. To circumven t this structural flaw and strictly preserve the relational ordering, we adopt a s equence-aware architecture. Specifically, given an ordered sequence of body rel ations [r1, r2,..., rl], an RNN [15] is employed to sequentially encode the path and generate the representation for the rule head $\mathrm{r_{l+1}}$. In detail, we perform a concatenation operation between each body relation embedding $\mathrm{r_i}$ and the com prehensive rule embedding $R$. The resulting fused representations serve as in puts to the RNN. To ensure that the network's output effectively aligns with t he actual rule head $\mathrm{r_{l+1}}$, we define the distance function as:

$$d(R,r_1,r_2,...,r_l) = \left\| RNN[R,r_1],[R,r_2],...,[R,r_l] - r_{l+1} \right\| \tag{4}$$

At this point, it replaces the relationship in the rule body or rule header wit h a random relationship. The loss function of logical rules is defined as:

$$L(R,\mathrm{r}_1,\mathrm{r}_2,...,\mathrm{r}_n) = -\log\sigma(\gamma_r - d(R,\mathrm{r}_1,\mathrm{r}_2,...,\mathrm{r}_n)) - \sum_{(R,\mathrm{r}_1^{'},\mathrm{r}_2^{'},...,\mathrm{r}_{l+1}^{'})\in M}\frac{1}{M}\log\sigma(d(R,\mathrm{r}_1^{'},\mathrm{r}_2^{'},...,\mathrm{r}_{l+1}^{'}) - \gamma_y) \quad (5)$$

**Joint Training Module:** The goal of the joint training module is to simulta neously optimize the representations of entities, relations, and logical rules with in a unified learning framework. Given a knowledge graph containing a set of factual triples K and a set of logical rules L, we jointly learn (i) entity–relati on embeddings through triple-level supervision (Eq.~2), and (ii) relation–rule e mbeddings through rule-level supervision (Eq.~4). Rather than optimizing these objectives independently, we integrate them into a single joint optimization pro cess. This design allows structural signals from logical rules to directly regulari ze entity–relation embeddings, while factual triples provide grounding constraint s for rule representations. As a result, the learned embeddings are encouraged t o be both structurally consistent with observed facts and semantically aligned with logical dependencies. Formally, the overall training objective is defined as:

$$L = \sum_{(h,r,t)\in k} L(h,r,t) + \alpha \sum_{(R,r_1,r_2,...,r_l)\in L} L(R,r_1,r_2,...,r_l) \quad (6)$$

where L(h,r,t) denotes the loss for factual triples, L (R, r_1, r_2, ..., r_l) denot es the logical rule loss, and \alpha is a hyperparameter that balances the relati ve importance of rule-based supervision. By jointly minimizing this objective, t he model effectively propagates rule-level constraints into the embedding space, which is particularly beneficial in sparse and low-resource settings where fact ual triples alone provide insufficient supervision.

**Selection:** After the joint optimization stage is completed, the pre-extractor will determine the adjacent entities that have the highest correlation with the f ew sample observation results based on the credibility scores of each entity. S pecifically, for a link prediction query (h, r, ?), we replace the missing tail wi th each candidate entity to form candidate triples (h, r, $t_{[1,2...,n]}$). Then, for each constructed triple, a knowledge graph embedding (KGE) score assessment is c onducted according to the following formula:

$$S = \left[S_1(h_1,r_1,t_{a1},...,t_{an}), S_2(h_2,r_2,t_{b1},...,t_{bn}),...,S_3(h_3,r_3,t_{c1},...,t_{cn})\right] \quad (7)$$

where k denotes the number of few-shot queries and n is the number of select ed candidate entities per query. By restricting the candidate space to the top-n high-scoring entities, this selection mechanism ensures that the expanded neighb ors are both structurally plausible and contextually consistent, thereby reducing noise propagation and enabling controllable downstream knowledge augmentatio n.

### 3.2 LLM-based Knowledge Fusion

This section describes the LLM-based knowledge fusion module, which perf orms controlled generative reasoning to enrich the knowledge graph using the high-confidence candidates produced by the pre-extractor. As illustrated in Fig.3. This module is designed to leverage the reasoning capability [29, 30] of LLM while strictly constraining their input space and output format to avoid uncont rolled hallucination.

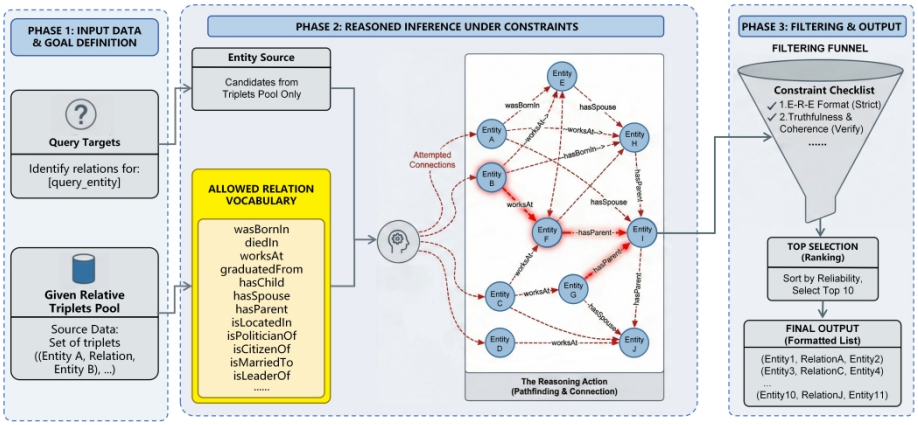


**Fig. 3.** The overview of LLM-based knowledge fusion.

Given a query (h, r, t), the pre-extractor first provides a compact candidate set T = $t_1$, $t_2$, ..., $t_n$} of plausible tail entities. Instead of prompting the LLM to freely explore the entire entity space, we construct structured prompt templates that explicitly specify the head entity h, relation r, and the candidate tails in T. The LLM is instructed to infer valid triples (h, r, t) by reasoning over the provided candidates and existing relational context in the dataset. Compared with CoT prompting strategies commonly used in prior work, our approach avoids multi-step open-ended reasoning that introduces excessive token overhead and unstable inference behavior. While CoT prompting decomposes a query into long reasoning traces, it often leads to increased computational cost and degraded performance for small or resource-constrained LLMs. In contrast, by supplying a pre-filtered candidate space, our method requires only a single, compact inference step, significantly improving efficiency and robustness.

To further ensure the reliability of knowledge augmentation, the prompt is explicitly designed to guide the Large Language Model in evaluating candidate triples along two critical dimensions: semantic consistency and factual plausibility. This includes a specific focus on temporal validity, ensuring that temporal attributes align with known event timelines and structural alignment with the relational topology of the existing knowledge graph. Formally, the LLM is instructed to select an optimal subset of candidate triples, denoted as $T^* \subseteq T$, such that each triple (h, r, t) $\subseteq$ $T^*$ satisfies contextual coherence with the input text and adheres to relational constraints derived from the graph's ontological schema. This process effectively embeds

generative flexibility within a structured knowledge boundary, mitigating the risk of semantic drift or factual contradiction.

Subsequently, a post-hoc filtering mechanism is applied, leveraging the pre-trained scoring function of the pre-extractor module to eliminate triples with low confidence scores or logical inconsistencies. This step yields a refined set of high-quality augmented triples that demonstrate both local compatibility and global graph consistency. These LLM-enhanced triples are then integrated into the training corpus to enrich the representation of entities and relations [24]. Following augmentation, multiple knowledge graph reasoning models, such as graph neural networks and symbolic rule-based systems, are deployed for downstream inference.

# 4 Experiments

## 4.1 Datasets:

To evaluate our approach, we conduct link prediction experiments on three benchmark datasets: YAGO3-10, FB15k-237 [12], and WN18RR [7]. We benchmark our enhanced knowledge graphs against their original counterparts utilizing two text-rich KGC baselines: SimKGC [21] and CSProm-KG [4]. These models were carefully selected based on their state-of-the-art performance, architectural novelty, and manageable computational overhead. Given the exhaustive candidate ranking required in KGC tasks, we primarily adopt baselines compatible with the highly efficient 1-N scoring paradigm. Furthermore, our evaluation incorporates representative structure-based embedding methods, specifically DistMult, RotatE, ComplEx, and HousE.

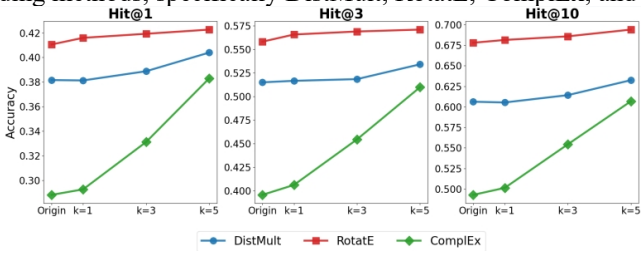


**Fig. 4.** Performance comparison on YAGO3-10 under different K-shot settings.

**Implementation Details:** The experimental operating system is Linux, the pr ocessor is Intel Xeon (R) Platinum 8457C CPU, the GPU is NVIDIA L20, the video memory is 48GB, and the CUDA version is 12.4. We evaluated SAGE on YAGO3-10, which is the standard benchmark for knowledge graph comple tion. We define few-shot entities as entities that appear in triplets less than or equal to K , where our K in {3,5,8} appears on the yago dataset. We compar ed our enhanced KG with the original KG using four structure-based KGC mo dels: DistMult[25], ComplEx[19], RotatE[16], and HousE[10]. As the LLM ba ckbone, we use Qwen-Plus[1] , with temperature set to 0.7, max generation le ngth 8192, and SEED=1234 for reproducibility. All baselines are implemented using their official code, with hyperparameters set to original defaults. We con

sider two variants: (1) using the pre-extractor to select top-30 candidate entities, and (2) using two-hop neighbors for LLM prompting.

### 4.2 Main Results:

Table1 and Fig. 4 report the predictive performance of YAGO3-10 under low shot settings. Overall, SAGE consistently outperforms both the original models and the two-hop augmentation baseline across different embedding architectures, achieving improvements in MRR as well as Hits@1, Hits@3, and Hits@10. For DistMult and RotatE, SAGE yields steady gains across all metrics, indicating that the proposed enhancement effectively complements conventional embedding models by enriching sparse relational contexts. In particular, the improvements in Hits@1 suggest that SAGE not only increases recall but also enhances the precision of top-ranked predictions. Notably, SAGE brings substantial improvements for ComplEx, where MRR and Hits@1 increase by a large margin over both the original and two-hop variants. This observation suggests that SAGE is particularly effective for models that are more sensitive to data sparsity, as the augmented knowledge helps alleviate insufficient structural supervision in long-tail regions.

**Table1**. Pediction results on YAGO3-10.

| Method / Metric | MMR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| **DistMult** | | | | |
| Origin_Model | 0.461 | 0.381 | 0.515 | 0.606 |
| Two-hop | 0.458 | 0.337 | 0.509 | 0.608 |
| SAGE | **0.484** | **0.404** | **0.534** | **0.632** |
| **RotatE** | | | | |
| Origin_Model | 0.503 | 0.410 | 0.538 | 0.678 |
| Two-hop | 0.508 | 0.412 | 0.564 | 0.686 |
| SAGE | **0.516** | **0.423** | **0.571** | **0.694** |
| **ComplEx** | | | | |
| Origin_Model | 0.359 | 0.288 | 0.395 | 0.492 |
| Two-hop | 0.381 | 0.307 | 0.421 | 0.522 |
| SAGE | **0.461** | **0.383** | **0.510** | **0.606** |
| **HousE** | | | | |
| Origin_Model | 0.570 | 0.491 | 0.618 | 0.713 |
| Two-hop | 0.568 | 0.489 | 0.614 | 0.715 |
| SAGE | **0.574** | **0.498** | **0.618** | **0.717** |

Table 2 reveals a nuanced performance pattern for SAGE on the WN18RR dataset. While SAGE consistently outperforms the base models in all metrics for DistMult and ComplEx, its gains are more marginal on RotatE and HousE, even showing slight performance regressions in specific metrics. For semantic-oriented models like DistMult and ComplEx, which rely primarily on bilinear transformations or complex-

space correlation operators, the contextual information injected by SAGE effectively compensates for the intrinsic data sparsity of WN18RR (with an average triple frequency of 4.28 and a long-tail entity proportion of approximately 30\%). By semantically enriching low-frequency entities and inferring plausible relational connections, SAGE enhances these models' capacity to capture associative latent features, thereby improving generalization, especially in sparse regions of the knowledge graph. In contrast, geometrically-constrained models such as RotatE and HousE explicitly encode relational patterns as rotations or other geometric transformations in vector space.

**Table2.** Prediction results on WN18RR.

| Method / Metric | MMR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| **DistMult** | | | | |
| Origin_Model | 0.431 | 0.391 | 0.445 | 0496 |
| SAGE | **0.441** | **0.496** | **0.453** | **0.536** |
| **RotatE** | | | | |
| Origin_Model | **0.473** | 0.425 | **0.491** | **0.571** |
| SAGE | 0.467 | **0.431** | 0.484 | 0.540 |
| **ComplEx** | | | | |
| Origin_Model | 0.440 | 0.415 | 0.466 | 0.512 |
| SAGE | **0.467** | **0.428** | **0.488** | **0.563** |
| **HousE** | | | | |
| Origin_Model | **0.511** | **0.465** | **0.528** | **0.602** |
| SAGE | 0.503 | 0.454 | 0.524 | 0.597 |

Table 3 reports the prediction performance of two descriptive knowledge graph completion methods on FB15K-237. SAGE consistently improves the performance of both descriptive models across all evaluation metrics. For SimKGC, SAGE increases MRR from 0.327 to 0.337 and yields gains of 1.0, 0.9, and 1.6 points on Hits@1, Hits@3, and Hits@10, respectively. These improvements indicate that augmenting the graph with structure-aware generated facts effectively compensates for insufficient structural signals that cannot be fully captured by textual descriptions alone. This suggests that SAGE provides complementary benefits even when rich semantic representations are available, by further enriching sparse relational structures rather than redundantly amplifying textual information.

**Table3.** Prediction results on FB15K-237.

| Method / Metric | MMR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| **SimKGC** | | | | |
| Origin_Model | 0.327 | 0.241 | 0.354 | 0496 |
| SAGE | **0.337** | **0.251** | **0.363** | **0.512** |
| **CSProm-KG** | | | | |
| Origin_Model | 0.352 | 0.261 | 0.387 | 0.536 |
| **SAGE** | **0.355** | **0.263** | **0.390** | **0.539** |

### 4.3 Ablation Studies

Table 4 presents the ablation results on YAGO3-10, highlighting the role of the Pre-extractor in SAGE. Compared with the original models, incorporating the Pre-extractor leads to consistent performance improvements across all evaluation metrics. For DistMult, SAGE improves MRR from 0.431 to 0.441 and increases Hits@10 by 4.0 points (from 0.496 to 0.536), indicating a notable enhancement in ranking quality for long-tail predictions. Similar trends are observed for ComplEx, where the inclusion of the Pre-extractor yields a substantial gain in MRR (from 0.440 to 0.472) and Hits@10 (from 0.512 to 0.563). These quantitative gains suggest that the Pre-extractor functions as a semantic anchor rather than a simple filtering module. By constraining the LLM-generated candidates to be structurally and contextually consistent with the underlying graph, it effectively bridges the gap between the LLM's open-ended generation space and the structured embedding space of the knowledge graph. In the absence of such constraints, generated facts are more likely to become distributional outliers, which degrades the scoring reliability of embedding-based models.

**Table4.** Results of ablation experiments conducted on the YAGO3-10 dataset.

| Method / Metric | MMR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| **SimKGC** | | | | |
| Origin_Model | 0.431 | 0.391 | 0.445 | 0496 |
| SAGE | **0.441** | **0.396** | **0.453** | **0.536** |
| **CSProm-KG** | | | | |
| Origin_Model | 0.440 | 0.415 | 0.466 | 0.512 |
| **SAGE** | **0.472** | **0.428** | **0.488** | **0.563** |

## 5 Conclusion

We propose SAGE, a neuro-symbolic framework addressing data sparsity for long-tail entities in Knowledge Graph Completion (KGC). It features a logic-enhanced pre-extractor for candidate generation and an LLM-based module for constrained semantic reasoning. This dual-stage design generates enriched triples for few-shot entities and mitigates LLM hallucinations, ultimately boosting downstream performance. Experiments on three benchmarks show SAGE consistently outperforms baselines, particularly in long-tail reasoning. However, the joint optimization process within the current framework remains computationally intensive, requiring a substantial number of training steps, and the performance gains are relatively marginal on certain models and datasets. In future work, we plan to adopt a more lightweight pre-extractor architecture and explore the integration of advanced LLMs with superior zero-shot reasoning and generalization capabilities, aiming to concurrently optimize computational efficiency and inference accuracy.

## References


1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., et al. (2023). Qwen technical report. arXiv Preprint. arXiv:2309.16609.
2. Bordes, A., Chopra, S., & Weston, J. (2014). Question Answering with Subgraph Embeddings. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 615–620
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems (NIPS), 26
4. Chen, C., Wang, Y., Sun, A., Li, B., & Lam, K.-Y. (2023). Dipping PLMs Sauce: Bridging Structure and Text for Effective Knowledge Graph Completion via Conditional Soft Prompting. Findings of the Association for Computational Linguistics: ACL 2023, 11489–11503
5. Chen, S., Cheng, H., Liu, X., Jiao, J., Ji, Y., & Gao, J. (2023). Pre-training transformers for knowledge graph completion. arXiv Preprint. arXiv:2303.15682
6. Cheng, K., Ahmed, N. K., & Sun, Y. (2023). Neural Compositional Rule Learning for Knowledge Graph Reasoning. Proceedings of the 11th International Conference on Learning Representations (ICLR)
7. Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2D Knowledge Graph Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1)
8. Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. Proceedings of the 22nd International Conference on World Wide Web, 413–422
9. Ge, X., Wang, Y.-C., Wang, B., & Kuo, C.-C. J. (2024). Knowledge Graph Embedding with 3D Compound Geometric Transformations. APSIPA Transactions on Signal and Information Processing, 13(1)
10. Li, R., Zhao, J., Li, C., He, D., Wang, Y., Liu, Y., Sun, H., Wang, S., Deng, W., Shen, Y., Xie, X., & Zhang, Q. (2022). HousE: Knowledge Graph Embedding with Householder Parameterization. Proceedings of the 39th International Conference on Machine Learning (ICML), 12693–12705
11. Li, Y., Yu, K., Huang, X., & Zhang, Y. (2022). Learning Inter-Entity-Interaction for Few-Shot Knowledge Graph Completion. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7691–7700
12. Mruthyunjaya, V., Pezeshkpour, P., Hruschka, E., & Bhutani, N. (2023). Rethinking language models as symbolic knowledge graphs. arXiv Preprint. arXiv:2308.13676
13. Niu, G., Li, Y., Tang, C., Geng, R., Dai, J., Liu, Q., Wang, H., Sun, J., Huang, F., & Si, L. (2021). Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 213–222
14. Qi, K., Du, J., & Wan, H. (2023). Learning from both structural and textual knowledge for inductive knowledge graph completion. Advances in Neural Information Processing Systems (NeurIPS), 36, 26546–26562
15. Qu, M., Chen, J., Xhonneux, L.-P., Bengio, Y., & Tang, J. (2021). RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs. Proceedings of the 9th International Conference on Learning Representations (ICLR)

16. Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. Proceedings of the 7th International Conference on Learning Representations (ICLR)
17. Tang, X., Zhu, S.-C., Liang, Y., & Zhang, M. (2022). RulE: Neural-Symbolic Knowledge Graph Reasoning with Rule Embedding. Advances in Neural Information Processing Systems (NeurIPS), 35, 28830–28842
18. Tang, Y., Huang, J., Wang, G., He, X., & Zhou, B. (2020). Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2713–2722
19. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. Proceedings of the 33rd International Conference on Machine Learning (ICML), 2071–2080
20. Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., & Guo, M. (2018). RippleNet: Propagating user preferences on the knowledge graph for recommender systems. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 417–426
21. Wang, L., Zhao, W., Wei, Z., & Liu, J. (2022). SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 4281–4294
22. Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. Proceedings of the AAAI Conference on Artificial Intelligence, 28(1)
23. Xu, D., Xu, T., Wu, S., Zhou, J., & Chen, E. (2022). Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. Proceedings of the 30th ACM International Conference on Multimedia (MM '22), 3857–3866
24. Xu, D., Zhang, Z., Lin, Z., Wu, X., Zhu, Z., Xu, T., Zhao, X., Zheng, Y., & Chen, E. (2024). Multi-perspective Improvement of Knowledge Graph Completion with Large Language Models. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 11985–11996
25. Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2015). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. Proceedings of the 3rd International Conference on Learning Representations (ICLR)
26. Yu, L., Luo, Z., Liu, H., Lin, D., Li, H., & Deng, Y. (2022). TripleRE: Knowledge graph embeddings via tripled relation vectors. arXiv Preprint. arXiv:2209.08271
27. Zhang, C., Yao, H., Huang, C., Jiang, M., Li, Z., & Chawla, N. V. (2020). Few-shot knowledge graph completion. Proceedings of the AAAI Conference on Artificial Intelligence, 34(3), 3041–3048
28. Zhang, W., Paudel, B., Wang, L., Chen, J., Zhu, H., Zhang, W., Bernstein, A., & Chen, H. (2019). Iteratively learning embeddings and rules for knowledge graph reasoning. Proceedings of the World Wide Web Conference (WWW), 2366–2377
29. Yang Z, Yang Y, Zhao C, et al. (2024). Perllm: Personalized inference scheduling with edge-cloud collaboration for diverse llm services. arXiv preprint arXiv:2405.14636, 2024.
30. Tian Y, Yang Z. (2025). SAEC: Scene-Aware Enhanced Edge-Cloud Collaborative Industrial Vision Inspection with Multimodal LLM. arXiv preprint arXiv:2509.17136